IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of	)
	)
MORIMOTO et al.	)
	)
Application Number: To be Assigned	)
	)
Filed: Concurrently Herewith	)
	)
For: CROSS LINGUAL TEXT CLASSIFICATION APPARATUS AND METHOD	)
	)
ATTORNEY DOCKET NO. ASAM.0113	)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of September 29, 2003, the filing date of the corresponding Japanese patent application 2003-338177.

A certified copy of Japanese patent application 2003-338177 is being submitted herewith.

Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344

Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
February 24, 2004

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２００３年　９月２９日

出 願 番 号
Application Number: 特願２００３－３３８１７７

[ST. 10/C]: [ＪＰ２００３－３３８１７７]

出 願 人
Applicant(s): 株式会社日立製作所

２００４年　２月　９日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００４－３００７８７９

【書類名】　特許願
【整理番号】　GM0309083
【提出日】　平成15年 9月29日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/27
【発明者】
　【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所中央研究所内
　【氏名】　森本　康嗣
【発明者】
　【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所中央研究所内
　【氏名】　梶　博行
【発明者】
　【住所又は居所】　東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所中央研究所内
　【氏名】　今一　修
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100075513
　【弁理士】
　【氏名又は名称】　後藤　政喜
【選任した代理人】
　【識別番号】　100084537
　【弁理士】
　【氏名又は名称】　松田　嘉夫
【選任した代理人】
　【識別番号】　100114236
　【弁理士】
　【氏名又は名称】　藤井　正弘
【手数料の表示】
　【予納台帳番号】　019839
　【納付金額】　21,000円
【その他】　国等の委託研究の成果に係る特許出願（平成１４年度新エネルギー・産業技術総合開発機構基盤技術研究促進事業委託研究、産業活力再生特別措置法第３０条の適用を受けるもの）
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0110326

【書類名】特許請求の範囲
【請求項１】
文書の入力を受け付ける文書入力部と、入力された分類対象文書の分類に用いられる概念シソーラスファイルと、第１及び第２の言語を含む複数言語に対応する複数言語語義分類知識ファイルと、単語分類知識ファイルとを記憶する記憶部と、上記入力された分類対象文書の分類を実行してカテゴリ付与を行う処理装置と、上記カテゴリ付与結果を出力する出力部とを有する文書分類装置であって、
上記文書入力部は、上記第１の言語の分類対象文書の入力を受け付け、
上記処理装置は、
上記第１の分類対象文書から単語を抽出し、
上記概念シソーラスファイルを用いて上記抽出された単語を語義に変換し、
上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と該変換された語義とを照合して各カテゴリについての第１のスコアを計算し、
上記単語分類知識ファイルに含まれる単語分類情報と上記抽出された単語とを照合して各カテゴリについての第２のスコアを計算し、
各カテゴリについての上記第１及び第２のスコアを統合して上記第１の言語の分類対象文書のカテゴリを決定してカテゴリ付与を行い、
上記単語分類知識ファイルは、上記第１の言語のカテゴリ情報付き文書に含まれる単語を用いて単語をベースとする単語分類知識を学習して生成されることを特徴とする文書分類装置。
【請求項２】
請求項１記載の文書分類装置であって、
該文書分類装置を用いてカテゴリを付与された分類対象文書は、上記単語分類知識ファイルの生成に用いられる上記第１の言語のカテゴリ情報付き文書として単語分類知識の学習に用いられることを特徴とする文書分類装置。
【請求項３】
請求項１記載の文書分類装置であって、
上記複数言語語義分類知識ファイルは、
上記第２の言語のカテゴリ情報付き文書に含まれる単語を抽出して上記概念シソーラスファイルを用いて語義に変換し、該語義と上記第２の言語のカテゴリ情報付き文書のカテゴリ情報とを用いて語義分類知識を学習して生成されることを特徴とする文書分類装置。
【請求項４】
請求項１記載の文書分類装置であって、表示装置とユーザ入力を受け付けるユーザ入力装置とを有し、
上記表示装置は、上記カテゴリ付与に用いられる語義を表示し、上記分類対象文書から抽出された一の単語に対する語義候補が複数存在する場合には上記分類付与に用いられる語義以外の語義候補をあわせて表示し、
ユーザ入力装置を介して上記複数の語義候補のいずれかを選択する入力を受け付け、
該入力された選択情報に基づいて上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と照合する語義を変更して上記カテゴリ付与を行うことを特徴とする文書分類装置。
【請求項５】
請求項１記載の文書分類装置であって、表示装置とユーザ入力を受け付けるユーザ入力装置とを有し、
上記処理装置は、上記分類対象文書に付与されたカテゴリについての上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる第１の語義と、上記分類対象文書に含まれる単語を上記概念シソーラスで変換して得られる第２の語義とを比較して矛盾する語義を有する単語を抽出し、
上記表示装置は、該抽出された単語に対する上記第１及び第２の語義を表示し、
ユーザ入力装置を介して上記第１及び第２の語義のいずれかを選択する入力を受け付け

該入力された選択情報に基づいて上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる各カテゴリに対応する語義を変更することを特徴とする文書分類装置。

【請求項６】

文書の入力を受け付ける文書入力部と、入力された分類対象文書のカテゴリに用いられる概念シソーラスファイルと、第１及び第２の言語を含む複数言語に対応する複数言語語義分類知識ファイルと、単語分類知識ファイルとを記憶する記憶部と、上記入力された分類対象文書の分類を実行してカテゴリ付与を行う処理装置と、上記カテゴリ付与結果を出力する出力部とを有する文書分類装置であって、

上記文書入力部は、上記第１の言語の分類対象文書の入力を受け付け、

上記処理装置は、

上記第１の分類対象文書から単語を抽出し、

上記概念シソーラスファイルを用いて上記抽出された単語を語義に変換し、

上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と該変換された語義とを照合して各カテゴリについての第１のスコアを計算し、

上記単語分類知識ファイルに含まれる単語分類情報と上記抽出された単語とを照合して各カテゴリについての第２のスコアを計算し、

各カテゴリについての上記第１及び第２のスコアを統合して上記第１の言語の分類対象文書のカテゴリを決定してカテゴリ付与を行い、

上記単語分類知識ファイルは、上記第１の言語の教師文書から服すの単語間の関係を示す情報を抽出し、該抽出された単語間の関係情報と上記複数言語語義分類知識ファイルに含まれる各カテゴリの単語分類情報とを用いて各カテゴリの単語分類知識を抽出して生成されることを特徴とする文書分類装置。

【請求項７】

請求項６記載の文書分類装置であって、

上記複数言語語義分類知識ファイルは、

上記第２の言語のカテゴリ情報付き文書に含まれる単語を抽出して上記概念シソーラスファイルを用いて語義に変換し、該語義と上記第２の言語のカテゴリ情報付き文書のカテゴリ情報とを用いて語義分類知識を学習して生成されることを特徴とする文書分類装置。

【請求項８】

請求項６記載の文書分類装置であって、表示装置とユーザ入力を受け付けるユーザ入力装置とを有し、

上記表示装置は、上記カテゴリ付与に用いられる語義を表示し、上記分類対象文書から抽出された一の単語に対する語義候補が複数存在する場合には上記カテゴリ付与に用いられる語義以外の語義候補をあわせて表示し、

ユーザ入力装置を介して上記複数の語義候補のいずれかを選択する入力を受け付け、

該入力された選択情報に基づいて上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と照合する語義を変更して上記カテゴリ付与を行うことを特徴とする文書分類装置。

【請求項９】

請求項６記載の文書分類装置であって、表示装置とユーザ入力を受け付けるユーザ入力装置とを有し、

上記処理装置は、上記分類対象文書に付与されたカテゴリについての上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる第１の語義と、上記分類対象文書に含まれる単語を上記概念シソーラスで変換して得られる第２の語義とを比較して矛盾する語義を有する単語を抽出し、

上記表示装置は、該抽出された単語に対する上記第１及び第２の語義を表示し、

ユーザ入力装置を介して上記第１及び第２の語義のいずれかを選択する入力を受け付け

該入力された選択情報に基づいて上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる各カテゴリに対応する語義を変更することを特徴とする文書分類装置。

【請求項１０】

入力された分類対象文書の分類に用いられる概念シソーラスファイルと、第１及び第２の言語を含む複数言語に対応する複数言語語義分類知識ファイルと、単語分類知識ファイルとを記憶する記憶部と、上記入力された分類対象文書の分類を実行してカテゴリ付与を行う処理装置とを有する文書分類装置において入力された分類対象文書にカテゴリを付与する文書分類方法であって、

上記第１の言語の分類対象文書の入力を受け付け、

上記第１の分類対象文書から単語を抽出し、

上記概念シソーラスファイルを用いて上記抽出された単語を語義に変換し、

上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と該変換された語義とを照合して各カテゴリについての第１のスコアを計算し、

上記単語分類知識ファイルに含まれる単語分類情報と上記抽出された単語とを照合して各カテゴリについての第２のスコアを計算し、

各カテゴリについての上記第１及び第２のスコアを統合して上記第１の言語の分類対象文書のカテゴリを決定してカテゴリ付与を行い、

上記単語分類知識ファイルは、上記第１の言語のカテゴリ情報付き文書に含まれる単語を用いて単語をベースとする単語分類知識を学習して生成されることを特徴とする文書分類方法。

【請求項１１】

請求項１０記載の文書分類方法であって、

該文書分類装置を用いてカテゴリを付与された分類対象文書は、上記単語分類知識ファイルの生成に用いられる上記第１の言語のカテゴリ情報付き文書として単語分類知識の学習に用いられることを特徴とする文書分類方法。

【請求項１２】

請求項１０記載の文書分類方法であって、

上記第２の言語のカテゴリ情報付き文書に含まれる単語を抽出して上記概念シソーラスファイルを用いて語義に変換し、該語義と上記第２の言語のカテゴリ情報付き文書のカテゴリ情報とを用いて語義分類知識を学習して生成されることを特徴とする文書分類方法。

【請求項１３】

請求項１０記載の文書分類方法であって、上記文書分類装置はさらに表示装置とユーザ入力装置とを有し、

上記表示装置に、上記カテゴリ付与に用いられる語義を表示し、

上記分類対象文書から抽出された一の単語に対する語義候補が複数存在する場合には上記カテゴリ付与に用いられる語義以外の語義候補をあわせて表示し、

ユーザ入力装置を介して上記複数の語義候補のいずれかを選択する入力を受け付け、

該入力された選択情報に基づいて上記複数言語語義分類知識ファイルに含まれる各カテゴリの情報と照合する語義を変更して上記カテゴリ付与を行うことを特徴とする文書分類方法。

【請求項１４】

請求項１０記載の文書分類方法であって、上記文書分類装置はさらに表示装置とユーザ入力装置とを有し、

上記分類対象文書に付与されたカテゴリについての上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる第１の語義と、上記分類対象文書に含まれる単語を上記概念シソーラスで変換して得られる第２の語義とを比較して矛盾する語義を有する単語を抽出し、

該抽出された単語に対する上記第１及び第２の語義を上記表示装置に表示し、

ユーザ入力装置を介して上記第１及び第２の語義のいずれかを選択する入力を受け付け

該入力された選択情報に基づいて上記複数言語語義分類知識ファイルまたは上記単語分類知識ファイルに含まれる各カテゴリに対応する語義を変更することを特徴とする文書分類方法。

【書類名】明細書
【発明の名称】複数言語を対象とした文書分類装置及び文書分類方法
【技術分野】
【０００１】
本発明は、人手によってカテゴリが付与された文書から文書を分類するための知識を学習し、学習された知識を用いて文書を分類する文書分類装置及び文書分類方法に関し、特に、複数言語の文書を対象として文書を分類する言語横断的なものに関する。
【背景技術】
【０００２】
ワープロやＰＣなどの普及により、ほとんどの文書が電子的に作成されるようになったことから、計算機上で扱うことができる電子化文書の量が増大している。このような状況に対処するための一つの技術として自動文書分類技術が開発されている。自動文書分類技術は、人手によってカテゴリが付与された文書から学習された文書を分類するための知識を用いて、カテゴリが付与されていない文書に対し新たにカテゴリを付与する技術である。
【０００３】
従来、文書分類技術は、例えば日本語のようなある１つの言語において分類を行うために利用されてきたが、現在、インターネットの普及やグローバル化の進展に伴って、複数言語の文書を扱うニーズが大きくなっている。このような目的に対応するため、例えば、特開平９－６７９９号公報（文書分類装置及び文書検索装置）に開示されているように、人手で作成した概念辞書を用いて特定の言語とは独立に文書分類を行う技術する技術が存在する。また、別のアプローチとして、特開２００３－７６７１０号公報（多言語情報検索システム)に開示されているような、対訳辞書を用いて検索条件を翻訳した後、文書を検索する技術などが存在する。特開２００３－７６７１０号公報に開示されている技術は、文書検索に関するものであるが、同様の考え方を文書分類に適用することが可能である。すなわち、文書を翻訳し、ある言語、例えば英語に統一してから処理を行うなどの方法により、文書分類に適用できる。
【特許文献１】特開平９－６７９９号公報
【特許文献２】特開２００３－７６７１０号公報
【発明の開示】
【発明が解決しようとする課題】
【０００４】
しかしながら、従来の技術には以下のような問題が存在する。
(1)人手作成による概念辞書に基づく方式
人手作成の概念辞書に基づく方式の場合、概念辞書を構築することが極めて困難であり、コストが非常に大きくなるため、現実的にはシステムの構築が困難である。特に対象分野が広い場合には、概念辞書を人手で作成することは、非常に困難である。
(2)対訳辞書を用いた翻訳に基づく方式
翻訳に基づく方式の場合、対訳辞書が十分に整備されていない場合には、訳語が得られないため分類精度が低下するという問題がある。対訳辞書の網羅性を上げるためには、概念辞書の場合と同様に大きなコストがかかるため、対訳辞書の網羅性が高いことを前提とした技術は現実的ではない。
【０００５】
また、対訳辞書には、一つの見出し語に対して複数の訳語が存在するという曖昧性の問題が存在する。通常、翻訳方式では、機械翻訳の技術により訳語を１個選択するか、あるいは、存在する訳語全てに翻訳する手法がとられるが、前者の場合、漏れが発生し易くなり、後者の場合、ノイズが発生し易くなる。
【０００６】
本発明の第１の課題は、概念、あるいは語義をベースとして文書の分類を行う複数言語を対象とした分類方式を提供することにある。その際、従来技術とは異なり、自動学習し

た知識を用いて、ある言語の単語を言語独立な概念に変換する技術に基づいて、概念ベースを自動構築することによって、現実的なコストで複数言語を対象とした分類システムを構築できるための方法を提供する。

【０００７】

本発明の第２の課題は、対訳情報に不備が存在する場合でも、従来技術より正確な分類結果が得られる複数言語を対象とした分類方式を提供することにある。

【０００８】

本発明の第３の課題は、複数言語のカテゴリ付与済み文書が存在する場合に、個々の言語において従来技術により分類システムを構成した場合と比較して、より正確な分類結果が得られる複数言語を対象とした分類方式を提供することにある。

【０００９】

本発明の第４の課題は、語義を用いることにより、インタラクティブに分類結果や分類知識を修正する方法を提供することにある。

【課題を解決するための手段】

【００１０】

本発明の第１の課題は、カテゴリが付与されていない少なくとも２種類の言語の文書集合から単語を語義に変換するための知識を学習し、カテゴリが付与されている第１の言語の文書から抽出される単語の集合を前記単語を語義に変換するための知識を用いてカテゴリ毎の語義の集合に変換し、

前記カテゴリ毎の語義の集合から語義からなる分類知識を学習し、

分類対象である第２の言語の文書から抽出される単語の集合を語義の集合に変換し、

前記第２の言語の文書から抽出された語義の集合と、語義からなる分類知識を比較して、前記第２の言語の文書のカテゴリを決定することにより解決できる。

【００１１】

本発明の第２の課題は、上記に加え、

カテゴリが付与されていない第２の言語の文書から単語間の共起情報を抽出し、

前記学習された語義からなる分類知識を構成する第２の言語の単語を取得し、

前記取得された第２の単語と関連が強い単語を前記共起情報に基づいて抽出し、

抽出された単語を第２の言語のための分類知識として用い、前記分類対象である第２の言語の文書から抽出される単語の集合とを比較して、スコアを計算し、

前記単語に基づくスコアと、前記語義によるスコアとを統合することにより決定されるスコアに基づいて、前記分類対象である第２の言語の文書のカテゴリを決定することにより解決できる。

【００１２】

本発明の第３の課題は、第１の課題を解決する手段に加え、

カテゴリが付与されている第２の言語の文書から単語をベースとする分類知識を学習し、

前記単語に基づく分類知識と、前記分類対象である第の言語の文書から抽出される単語の集合とを比較して、スコアを計算し

前記単語に基づくスコアと、前記語義によるスコアとを統合することにより決定されるスコアに基づいて、前記分類対象である第２の言語の文書のカテゴリを決定することにより解決できる。

【００１３】

本発明の第４の課題は、第１の課題を解決する手段に加え、

複数の語義を持つ単語について可能な語義を表示し、ユーザに選択させ、

ユーザが選択した語義に基づいて、スコアを再計算し、カテゴリを再決定し、

最終的に決定されたカテゴリの分類知識と前記処理対象文書から抽出される語義の集合とを比較して矛盾する語義を持つ単語を検出し、

前記検出された単語をユーザに表示し、語義を選択させ、

選択された語義にしたがって、語義分類知識を修正する

ことにより解決できる。

【発明の効果】

【００１４】

本発明によれば、日本語と英語のような複数言語の文書を分類するシステムを現実的なコストで実現することが可能となる。また、本発明による文書分類システムでは、対訳情報が不備である場合にも、従来技術と比較して高精度な文書分類が可能である。さらに、分類の対象となる複数の言語のそれぞれにおいて、従来技術による分類システムを構築した場合と比較して、より高い精度を持つ分類システムを構築できる。さらに、語義を利用することによって、人間に分かりやすい方法で、インタラクティブに分類結果や分類知識を修正することが可能になる。

【発明を実施するための最良の形態】

【００１５】

以下、本発明の実施例を、図面を用いて説明する。

【００１６】

図１に、本発明による２言語文書分類支援システムのブロック図を示す。本発明による２言語文書分類支援システムは、文書にカテゴリを付与するための分類知識を学習するプログラム群と学習された分類知識を用いてカテゴリを付与するためのプログラム群からなる。ここでは、例として日本語と英語の２言語に対応する文書分類支援システムを説明するが、言語はこの２言語に限られるものではない。また、対応する言語が３言語以上であっても同様の手法で文書分類、学習等を行うことが可能である。

【００１７】

分類知識を学習するためのプログラム群は、日本語文書101および英語文書102と対訳辞書103から概念シソーラス104を生成する概念シソーラス生成プログラム１、カテゴリ付き日本語文書105およびカテゴリ付き英語文書106を単語に分割し、さらに単語を文書中の語義として正しい語義に変換する語義変換プログラム2、文書毎に語義に変換された単語のデータから語義分類知識107を学習する語義分類知識学習プログラム3、カテゴリ付き日本語文書105から高精度日本語単語分類知識108を学習する日本語単語分類知識学習プログラム4、カテゴリ付き英語文書106から高精度英語単語分類知識109を学習する英語分類知識学習プログラム5、語義分類知識107と日本語文書101から低精度日本語単語分類知識110を学習する教師なし日本語単語分類知識学習プログラム6、語義分類知識107と英語文書102から低精度英語単語分類知識111を学習する教師なし英語単語分類知識学習プログラム7、を含む。日本語文書101、英語文書102はカテゴリが付与されていないものでもよいので、以下、カテゴリなし日本語文書およびカテゴリなし英語文書と呼ぶ。

【００１８】

カテゴリを付与するためのプログラム群は、語義分類知識107、高精度日本語単語分類知識108および低精度日本語単語分類知識110を用いて、分類対象日本語文書112にカテゴリを付与する日本語カテゴリ付与プログラム8、語義分類知識107、高精度英語単語分類知識109および低精度英語単語分類知識111を用いて、分類対象英語文書113にカテゴリを付与する英語カテゴリ付与プログラム9からなる。

【００１９】

本発明を実現するための装置の構成図を図２に示す。本発明を実現するための装置は、概念シソーラスプログラム1などのプログラム類、カテゴリなし日本語文書101などのデータ・ファイル類を格納する記憶装置100、各種の処理を実行するＣＰＵ200、ユーザからの入力を受け付けるキーボードやマウス、処理結果をユーザに対して表示するためのディスプレイなどからなる入出力装置300からなる。なお、図２では、一例としてスタンドアロンシステムとしての構成が記載されているが、これに限定するものではない。例えば、分類知識を学習する装置とカテゴリを付与する装置を別々の装置として構成することも可能である。また、分類知識を学習する装置についても、分類知識の種類毎あるいはカテゴリ毎に別の装置として構成することも可能である。さらに、カテゴリを付与する装置を複数のユーザが同時に使用できるクライアント・サーバ型の構成とすることも可能である。

【００２０】
以下では、本実施例における日英２言語文書分類支援システムの処理について説明する。まず、図３の処理フローを用いて分類知識の学習処理について説明する。なお、以下の説明では、カテゴリなしの日本語および英語文書、カテゴリ付き日本語文書が存在する場合に、英語の文書を分類するシステムを構築する場合を想定して説明する。ただし、本システムは、図１の構成からも分かるように、日英の２言語に関して対象であり、上記の場合に限定するものではない。

【００２１】
カテゴリなし日本語文書およびカテゴリなし英語文書から対訳辞書を用いて概念シソーラスを生成する(ステップ11)。本ステップの処理は、特開２００１－３３１４８４号公報（パラレルシソーラスの生成プログラムを記録した記録媒体、パラレルシソーラスを記録した記録媒体及びパラレルシソーラスナビゲーションプログラムを記録した記録媒体)に開示されている技術を適用することによって実現することができるので詳細な説明は省略する。

【００２２】
概念シソーラスの概念図を図４に示す。概念シソーラスは、日本語と英語の単語集合の組で語義を表現し、日本語あるいは英語の単語が当該の語義を表しているかどうかを判断するための条件として、それぞれ日本語あるいは英語の単語の集合を持っている。語義を構成する日本語の単語、英語の単語をそれぞれ、日本語語義構成単語、英語語義構成単語と呼び、言語に依存しない総称として語義構成単語と呼ぶ。また、語義を判断するための条件となる単語を手がかり語という。例えば、図４に示された例の場合、「コート」という単語は、「テニス」のような単語とあわせて使用される場合は、[court,コート]という語義を持っているということを示している。関係のある語義間はリンクで連結されている。ただし、本実施例では、リンクの情報は特に利用しない。カテゴリが付与されていない文書は、現在ではインターネットなどを通じて、あるいは企業内などで大量に作成されているため、容易に収集することが可能である。

【００２３】
カテゴリ付き日本語文書とカテゴリ付き英語文書および概念シソーラスから、語義分類知識を学習する(ステップ12)。語義分類知識は、ステップ11で生成された概念シソーラス中の語義を単位として文書分類に必要な知識を記述している。そのため、分類の対象となる文書の言語には依存しない。また、語義分類知識学習の教師データとなるカテゴリ付き文書の言語にも依存しない。すなわち、例えば、カテゴリ付き日本語文書のみが存在し、カテゴリ付き英語文書が存在しない場合であっても、語義分類知識の学習は可能である。分類知識学習の教師データとなるカテゴリ付き文書を準備するコストは比較的高い。更に、例えば、日本語を母語とするユーザにとって、英語文書のカテゴリを判定するタスクは非常に負担となる。本発明では、一方の言語のカテゴリ付き文書から語義分類知識を学習できるため、ユーザの負担を軽減することが可能である。本ステップは、後で詳細に説明する。

【００２４】
カテゴリ付き英語文書から、高精度英語単語分類知識を学習する(ステップ13)。高精度英語単語分類知識は、英語単語を単位として文書分類に必要な知識を記述している。そのため、英語文書の分類にのみ使用される。また、カテゴリ、すなわち、正解が付与されたデータから学習されるため高精度である。本ステップは、後で詳細に説明する。

【００２５】
カテゴリ付き日本語文書から、高精度日本語単語分類知識を学習する(ステップ14)。本ステップは、対象が日本語であること以外は、ステップ13と同様に処理を行えるため、ステップ13の詳細な説明に準ずるものとし、説明は省略する。

【００２６】
カテゴリなし英語文書から、低精度英語単語分類知識を学習する(ステップ15)。低精度英語単語分類知識は、英語単語を単位として文書分類に必要な知識を記述している。その

ため、英語文書の分類にのみ使用される。また、カテゴリ、すなわち、正解がないデータから学習されるため低精度である。本ステップは、後で詳細に説明する。

【００２７】

カテゴリ付き日本語文書から、低精度日本語単語分類知識を学習する(ステップ16)。本ステップは、対象が日本語であること以外は、ステップ15と同様に処理を行えるため、ステップ15の詳細な説明に準ずるものとし、説明は省略する。

【００２８】

以上の処理により、語義分類知識、日本語単語分類知識、英語単語分類知識の3種類の分類知識が学習される。これらの分類知識を適宜組み合わせて使用することにより、より高い精度が得られる２言語分類技術を実現することができる。

【００２９】

以下では、ステップ12における語義分類知識の学習処理を、図５および図６を用いて詳細に説明する。図５は、語義分類知識の学習処理の概念図である。図６は、語義分類知識の学習処理の処理フローである。

【００３０】

カテゴリ付きの日本語文書あるいは英語文書を取り出し、これを単語に分割する(ステップ121)。文や文書を単語に分割するための形態素解析方法に関しては、例えば、特開２０００－２５９６２９号公報（形態素解析方法およびその装置）に開示されている方法を用いることができるため説明は省略する。単語分割結果を格納する単語テーブルの例を図７に示す。単語テーブルには、分割結果として得られた文書中に出現した単語が出現順序順に格納されている。

【００３１】

分割結果の単語を、概念シソーラスを用いて語義に変換する(ステップ122)。語義変換は、次のように行われる。語義に変換したい単語を注目語とし、注目語の近傍、例えば、前後Ｎ単語以内の語を文脈語とする。ただし、近傍の単語を文脈語として取得する場合には、対象となる品詞を名詞、動詞、形容詞などの内容語に限定する。対象とする品詞は予め定めておく。図７の例の場合、例えば、注目語を”コート”とし、前後2単語を文脈語とするものとすると、文脈語として、”シリコン”、”被覆する”、”材料”、”撥水性”が得られる。注目語について概念シソーラスを検索し、語義の候補と語義毎の手がかり語の集合を得る。手がかり語の集合と文脈語を照合し、最も妥当だと思われる語義を選択する。図４の例に示した概念シソーラスでは、”コート”を含む語義の手がかり語として、シリコン”、”撥水性”が得られる。これにより、語義として、[coat,コート,皮膜]が選択される。このような処理を、注目語を１個ずつずらしていくことにより、全ての単語について語義への変換を行う。

【００３２】

形態素解析結果を変換して得られた語義の系列を集計し、語義ベクトルを生成する(ステップ123)。語義ベクトルの例を図８に示す。語義ベクトルは、語義と文書中に出現した各語義の頻度からなる。

【００３３】

全てのカテゴリ付き文書を処理したかどうか調べる(ステップ124)。未処理の文書があればステップ121に戻り、全て処理済みであればステップ125に進む。

【００３４】

文書の語義ベクトルから語義分類知識を学習する(ステップ125)。分類知識の学習方法については、例えば、Lewis, D. D. and Ringuette, M., A comparison of two learning algorithms for text categorization, Third Annual Symposium on Document Analysis and Information Retrieval, 1994, pp.81-93 (以下、[Lewis])に記載の方法を用いることができるので詳細な説明は省略する。ここでは、一例として、Rocchio法による分類知識の学習方法について簡単に説明する。

【００３５】

まず、得られた語義ベクトルを統計処理し、語義特徴ベクトルを抽出する。Rocchio法

では、各文書の特徴ベクトルは、次の数式１のように定められる。

【００３６】

【数１】

$$\vec{d} = (d_1, d_2, \ldots)$$

$$d_j = \frac{f_j \times \log(m / m_j)}{\sqrt{\sum_k (f_k \times \log(m / m_k))^2}}$$

$\vec{d}$：文書dの特徴ベクトル

f_j：語義c_jの文書dにおける出現頻度（通常，語w_jの出現頻度）

m：学習用テキスト数

m_j：学習用テキスト中でc_jが出現した文書数

【００３７】

数式１の方法で計算される重みは、通常単語について計算されるためtf-idf値(Term Frequency － Inversed Document Frequency)と呼ばれている。

【００３８】

計算された語義特徴ベクトルと、元の文書に付与されたカテゴリの情報から語義分類知識を抽出する。カテゴリの特徴ベクトルは、各カテゴリに含まれる文書の語義特徴ベクトルの平均として、次の数式２のように定められる。

【００３９】

【数２】

$$\vec{C}_i = \frac{\sum_{d \in D_i} \vec{d}}{|D_i|}$$

$\vec{C}_i$：カテゴリC_iの特徴ベクトル

D_i：カテゴリC_iに含まれる文書の集合

【００４０】

語義分類知識の例を図９に示す。

【００４１】

以上の方法により、語義分類知識を学習することができる。なお、ここではRocchio法を一例として説明したが、必ずしもこの方法に限定するものではない。

【００４２】

以下では、ステップ13における高精度英語単語分類知識の学習処理を、図１０を用いて詳細に説明する。なお、高精度英語単語分類知識の学習処理は、図５で概念図を示した語義分類知識の学習処理とほぼ同等の処理によって実現できる。異なる点は、単語の語義への変換が行われるかどうかという点のみであり、それ以外は語義を単語に読み替えて理解すれば良い。

【００４３】

カテゴリ付きの英語文書を取り出し、これを単語に分割する(ステップ131)。これは、ステップ121と同様に処理を行うことが出来るため詳細な説明は省略するが、本実施例の想定では、ステップ121での処理が主に日本語が対象であったのに対し、本ステップでは英語が対象になる点について簡単に説明する。日本語では、単語の区切りが明確ではないのに対し、英語では空白文字を用いて単語の区切りを明確に示す。そのため、単語に分割処理は、日本語の場合と比較して容易である。そのため、ステップ121の説明で言及した特開２０００－２５９６２９号公報のような技術を用いず、より簡便に空白文字によって単語に分割した後、辞書を引いて活用形を標準形に変換する程度の処理を行うことが多い。分割結果である英語単語テーブルの例を図１１に示す。

【００４４】

形態素解析結果である単語の系列を集計し、英語単語ベクトルを生成する(ステップ132)。英語単語ベクトルの例を図１２に示す。英語単語ベクトルは、図８の語義ベクトルの語義を英語単語に置き換えた形式を持つ。

【００４５】

全てのカテゴリ付き文書を処理したかどうか調べる(ステップ133)。未処理の文書があればステップ131に戻り、全て処理済みであればステップ134に進む。

【００４６】

文書の英語単語ベクトルから英語単語分類知識を学習する(ステップ134)。これは、ステップ125と同様に処理を行うことができるため説明は省略する。

【００４７】

以下では、ステップ15における低精度英語単語分類知識の学習処理を、図１３を用いて詳細に説明する。

【００４８】

カテゴリなしの英語文書を取り出し、これを単語に分割する(ステップ151)。本ステップは、ステップ131と同様であり、分割結果である英語単語テーブルは図１１のようになる。

【００４９】

形態素解析結果から、単語の共起を抽出する(ステップ152)。単語の共起とは、「同時に」出現した単語の組である。共起を抽出する方法としては様々な方法が考えられるが、ここでは、単語列上でのウインドウを用いた方法について簡単に説明する。この方法は、ステップ122で説明した語義変換のための文脈語を取得する方法と類似している。図１１に示す英語単語テーブルを例として以下に説明する。単語テーブル上で、ある単語に注目し、これを注目語とする。注目語の近傍、例えば、前後Ｎ単語以内の語を共起語とする。ただし、共起語は、名詞、動詞、形容詞などの内容語に限定する。対象とする品詞は予め定めておく。図１１の例の場合、例えば、注目語を”coat”とし、前後2単語を共起語とすると、共起語として、”cover”、”silicon”、”water-shedding”、”material”が得られる。このとき、注目語と共起語の組を単語の共起として出力する。この例の場合、[coat,cover], [coat,silicon], [coat,water-shedding], [coat,material]が共起として抽出される。

【００５０】

教師なし学習によって、低精度英語単語分類知識を学習する(ステップ153)。以下、本ステップを詳細に説明する。

【００５１】

まず図１４に示した概念図を用いて基本的な考え方を説明する。語義分類知識学習の結果として、カテゴリAを特徴付ける語義として、[cell,電池]、[solar,太陽]が抽出され、カテゴリBを特徴付ける語義として、[cell,細胞]、[mitochondria,ミトコンドリア]が抽出されているとする。このとき、概念シソーラスの生成に用いる対訳辞書中に含まれない日英の単語の組は語義として抽出することができない。しかしながら、語義分類知識として抽出された語義と関連が強いにも関わらず、対訳が辞書に含まれていない語義が存在することは多くあると考えられる。これは、対訳辞書に含まれる語は、比較的一般性が高い語であり、一方、ある分野の専門用語などはカテゴリを判別するのに重要であるためである。図１４の例の場合、[amorphous,アモルファス]や[codon,コドン]という語義は、対訳辞書に含まれていないため抽出することができないが、各カテゴリをよく特徴付ける語義である。本発明では、単語間の関係情報に基づく教師なし学習により、このような語義を補足的に抽出する方法を提供する。すなわち、日英各言語において、単語間の関係の強さに関する情報を抽出しておき、各カテゴリの語義分類知識として抽出された語義を構成する単語と関係が強い単語を、各カテゴリを特徴付けるタームとして追加することによって、対訳辞書の不備によるカテゴリの特徴単語の漏れを防ぐ。単語間の関係を用いて単語をグループにまとめていく技術は、単語のクラスタリングと呼ばれており、例えば、辻井潤一編、言語と計算４、確率的言語モデル、東京大学出版会、1999にその方法が述べられている。しかしながら、単語のクラスタリング技術だけでは、単語の意味的曖昧性のため正しいクラスタリング結果を得ることは難しい。特に、人が予め定めたカテゴリの体系と、計算機によるクラスタリング結果が一致することは期待できない。本発明では、この問題を解決するため、語義分類知識獲得の際に既に得られている各カテゴリを特徴付ける語義

構成単語と単語間の関係によるクラスタリング技術を組み合わせることによって各カテゴリを特徴付ける単語を補足的に抽出する。具体的なクラスタリングアルゴリズムとしては様々な方法が考えられるが、以下では、単語の共起関係を利用する簡便な方法について図１５に示す処理フローにしたがって詳細に説明する。

【００５２】

ステップ151で抽出したカテゴリなし文書から単語のリストから１個の単語を取り出す(ステップ1531)。

【００５３】

取り出した単語と各カテゴリの語義分類知識の語義構成単語との強さを計算する(ステップ1532)。取り出した単語と各カテゴリの語義構成単語との組が共起であるかどうかを、ステップ152で抽出した共起のデータと比較する。共起であった場合には、その共起頻度の合計をカテゴリ毎に集計する。

【００５４】

得られたカテゴリ毎の共起頻度の合計の全カテゴリでの合計に対する割合を、取り出した単語と各カテゴリとの関連の強さとし、各カテゴリの低精度単語分類知識に格納する(ステップ1533)。ただし、頻度そのものが小さい場合には、統計的な信頼性が低いため、統計的な検定などにより信頼性を評価した後に割合を計算しても良い。あるいは、より簡便な方法としては、予め定めた閾値より頻度が小さい共起は除外する方法なども用いることができる。

【００５５】

以上では、共起を用いる簡便な方法を述べたが、これに限定するものではない。語義構成単語をあたかも種のように扱ってクラスタリングを行うことができる手法であれば、どのような手法でも適用することが可能である。例えば、Duda, R. O.,Hart, P. E., Stork, D. G., Pattern Classification, カナダ, Wiley-Interscience, 2002,pp.526-528に記載されているk-means法などを用いて実現することも可能である。

【００５６】

次に、図１６および図１７を用いてカテゴリの付与処理について説明する。図１６は、語義分類知識を用いたカテゴリ付与処理の概念図である。図１７は、カテゴリ付与処理全体の処理フローである。本実施例では、英文の文書が入力された場合を例に説明するが、日本語文書が入力された場合にも同様の処理を行うことができる。

【００５７】

処理対象文書を単語に分割し、得られた単語の系列を集計して、単語ベクトルを生成する(ステップ21)。本ステップはステップ121、ステップ132と同様に処理を行えば良いため、説明は省略する。

【００５８】

分割結果の単語を、概念シソーラスを用いて語義に変換し、得られた語義の系列を集計して、語義ベクトルを生成する(ステップ22)。本ステップはステップ122、ステップ123と同様に処理を行えば良いため説明は省略する。

【００５９】

単語ベクトルと低精度英語単語分類知識を照合し、各カテゴリに対するスコアを計算する(ステップ23)。この処理は、例えば、[Lewis]に記載の方法を用いることができるので詳細な説明は省略する。ここでは、一例として、Rocchio法によるカテゴリ付与方法について簡単に説明する。

【００６０】

Roccho法では、単語ベクトルをステップ125で説明したのと同様の方法で単語特徴ベクトルに変換し、単語特徴ベクトルと、語義分類知識を照合することによってスコアを計算する。具体的には、文書とカテゴリの間のスコアを次の数式３のように定める。

【００６１】

【数３】

$$sim_R(C_i, d) = \frac{\vec{C}_i \bullet \vec{d}}{\left|\vec{C}_i\right|}$$

sim_R：スコア

$\vec{C}_i$：カテゴリC_iの特徴ベクトル

$\vec{d}$：カテゴリC_iに含まれる文書の集合

【００６２】

単語ベクトルと高精度英語単語分類知識を照合し、各カテゴリに対するスコアを計算する(ステップ24)。

【００６３】

語義ベクトルと語義分類知識を照合し、各カテゴリに対するスコアを計算する(ステップ25)。本ステップは、単語を語義に読み替えることにより、ステップ25と同様に行えるので説明は省略する。

【００６４】

３種類のスコアを統合して総合的なスコアを計算し、このスコアに基づいて付与すべきカテゴリを決定する(ステップ26)。あるいは、このうち２種類のスコアを統合してカテゴリを決定してもよい。ここでは、一例としてカテゴリ付き英語文書を教師データとして用いる方法を説明するが、この方法に限定されるものではない。

【００６５】

総合的なスコアtsを以下の式で示すものとする。

【００６６】

ts = a*ws1 + b*ws2 + (1-a-b)*ms

ここで、ws1は低精度英語単語分類知識に基づくスコア、ws2は高精度英語単語分類知識に基づくスコア、msは語義分類知識に基づくスコアであり、a,bは0≦a≦1，0≦b≦1のパラメータである。パラメータは、例えば、以下のような方法により予め定めておく。カテゴリ付き英語文書に対し、ステップ21からステップ27の処理を実施し、ws1、ws2、msを求めておく。次に、a,bを例えば、0から0.05刻みに変化させ、仮のtsを計算し、このtsに基づいてカテゴリ付与を行い、付与済みのカテゴリと比較して正解であるかどうかを評価する。これを全ての、またはカテゴリ付与の信頼性を評価するに十分な数のカテゴリ付き英語文書（すなわち、分類付与の正解がわかっている文書）に対して実施することで正解率

を求めることができる。最終的に、正解率が最も高いa,bをパラメータとして用いる。

【００６７】

本発明では、分類知識を学習する３種類の方法、すなわち、教師あり単語分類知識学習、語義分類知識学習、教師なし単語分類知識学習を利用しているが、教師あり単語分類知識学習が最も分類精度が良い分類知識を学習可能であり、教師なし単語分類知識学習による結果が最も分類精度が悪いと考えられる。また、各学習方法が利用できるデータの量の多寡が学習された分類知識による分類精度に影響する。すなわち、データ量が多いほど、分類精度が向上する。本発明の目的は、任意の時点で利用可能なデータを最大限に活用した、高い精度の分類結果を得ることであり、以上のようにスコアを定めることにより、この目的を達成することができる。例えば、このパラメータは一度決定されたら変更されないものではなく、各データ量の変化によって適宜変更することが望ましい。例えば、本実施例で想定した状況では、初期状態において教師ありデータとして利用できるカテゴリ付き英語文書は全く存在しないか、あっても少量である。そのため、語義分類知識学習あるいは教師なし単語分類知識学習によって学習された分類知識を利用してカテゴリ付与を行う。一方、システムが運用され時間が経過すると、システムが付与したカテゴリを人手によってチェックするといった形態で、カテゴリ付与済み文書の量が増大していく。カテゴリ付与済み文書の量がある程度増大した後では、教師あり単語分類知識学習によって学習された分類知識を重視する、すなわちパラメータ学習の結果bの値が大きくなることにより、分類精度が向上する。このように、任意の時点で利用可能なデータにシステムを最適化することが可能となる。なお、上記の方法だけでは、カテゴリ付き英語文書が全く存在しない場合には、パラメータを決定することができないが、このような場合でも、以下の代替案によってパラメータを決定することが可能である。代替案としては、カテゴリ付き日本語文書を利用する。カテゴリ付き日本語文書について、ステップ26と同様の方法によって、パラメータを決定する。この場合、bは0となり、aの値、すなわち、教師なし単語分類知識学習の結果をどの程度重視するかを決定することになる。この結果を用いて総合的なスコアを計算する。

【００６８】

インタラクティブにカテゴリ付与の誤りを修正する(ステップ27)。以下では、本ステップを詳細に説明する。

【００６９】

従来技術による単語に基づいた文書分類では、カテゴリ付与の誤りは、ある単語が出現した場合にどのカテゴリである確率が高いかに関する確率の推定の誤りだと考えられる。しかしながら、確率の推定値の正しさを人間が判断することはほぼ不可能である。一方、本発明では、単語の語義を導入しているため、単語を語義に変換する処理と語義に基づいてカテゴリを決定する処理での誤りを累積したものが全体の誤りとなる。ここで、語義が正しく選択されていれば、カテゴリを決定する処理での誤りが小さくなる。一方、単語の語義への変換処理の誤りは比較的人間が判断し易い形で提示することができるため、語義を導入することにより、インタラクティブにカテゴリを付与することができる文書分類システムを構築できる。

【００７０】

また、２言語文書分類システムでは、システムのユーザはいずれか一方の言語を母語とすると考えられる。本実施例の場合、日本語を母語とするユーザが、英語の文書を分類することを想定している。このような場合、英語の文書を理解し、その内容に応じて、付与されたカテゴリが正しいかどうかを判断することは難しい。本発明では、カテゴリ付与に使用された語義をユーザが母語とする言語(本実施例では日本語)で表示することによって、適切なカテゴリが付与されたかどうかの判断を支援することができる。

【００７１】

図１８は、インタラクティブなカテゴリ付与誤り修正画面の表示例を示す。ここでは、カテゴリを付与する際に用いられた語義および当該語義と別解になり得る語義の一覧が示されている。ある語義が別解となり得る語義を持つ場合には、ある語義と別解の語義が併

置して表示され、別解が存在しない場合には、当該語義のみが表示されている。別解が存在する場合には、解毎にチェックボックスが表示され、優先された解にチェックが付与される。図１８の例の場合、分類対象文書中に”cell”という単語が存在したため、｛cell,電池｝、｛cell,細胞｝という２個の語義が別解としてあり得る。よって、この２個の語義が併置されており、また、優先された解である、[cell,電池]にチェックが付与されている。また、”energy”という単語に対し、[energy,エネルギー]という語義が存在するが、”energy”に対しては語義の曖昧性が存在しないため、[energy,エネルギー]のみが表示されている。また、曖昧性がある単語”panel”に対しては、誤った語義[panel,回答者]が選択されている。このような表示をユーザが参照すると、[cell,電池]、[energy,エネルギー]、[silicon,シリコン,ケイ素]といった語義が表す文脈に対し、選択されている語義[panel,回答者]よりも[panel,パネル,はめ板]の方が語義としてふさわしいことが判断できる。よって、ユーザはチェックを[panel,パネル,はめ板]に対応するチェックボックスに付与しなおす。これにより、語義ベクトルの内容が変更され、同時に各カテゴリに対するスコアを再計算する。再計算は、ユーザが明示的に行うようにしても良い。以上のようにユーザが正しい語義を選択することによって、簡便な操作で分類精度を向上することができる。

【００７２】

インタラクティブに分類知識を修正する(ステップ28)。本ステップを詳細に説明する。

【００７３】

ステップ28の処理を含め、最終的に処理対象文書に付与するカテゴリが決定されたとする。このとき、付与するカテゴリの分類知識と処理対象文書の語義ベクトルを比較して、矛盾する語義を抽出する。矛盾する語義とは、同じ単語に対して複数の語義があり得る場合、その複数の語義同士のことをいう。図１９に例を示す。図１９の例の場合、処理対象文書中に、[panel,パネル,はめ板]という語義が含まれており、分類知識中には[panel,回答者]という語義が含まれていることを示している。この２つの語義は、”panel”の２種類の語義であり、同じカテゴリの文書であれば、同一の語義が使用される可能性が高いため、矛盾すると呼ぶ。このような場合は、図１９のように分類知識中の矛盾する語義を表示し、ユーザに確認を求める。ユーザが修正を求める入力を行った場合には、分類知識中の語義が変更される。以上の処理により、インタラクティブに分類知識を修正することが可能となる。

【００７４】

以上の処理により、入力された英語文書に対し、カテゴリを付与することができる。従来技術では、ステップ23の単語ベクトルと英語単語分類知識との照合によるスコアのみによりカテゴリが決定されていたのに対し、本実施例では日本語の文書から学習された語義分類知識との照合によるスコアも加味してスコアを決定することができるため、日英いずれかの言語の文書のみにより分類システムを構成した場合と比較して、より高い分類精度を達成することが可能となる。

【図面の簡単な説明】

【００７５】

【図１】本発明の実施例である日英２言語文書分類支援システムのブロック図である。

【図２】日英２言語文書分類支援システムの装置の構成図である。

【図３】日英２言語文書分類支援システムの処理フローである。

【図４】概念シソーラスの概念図である。

【図５】語義分類知識学習処理の概念図である。

【図６】語義分類知識学習処理の処理フローである。

【図７】日本語単語テーブルの例である。

【図８】語義ベクトルの例である。

【図９】語義分類知識の例である。

【図１０】高精度英語単語分類知識学習処理の処理フローである。

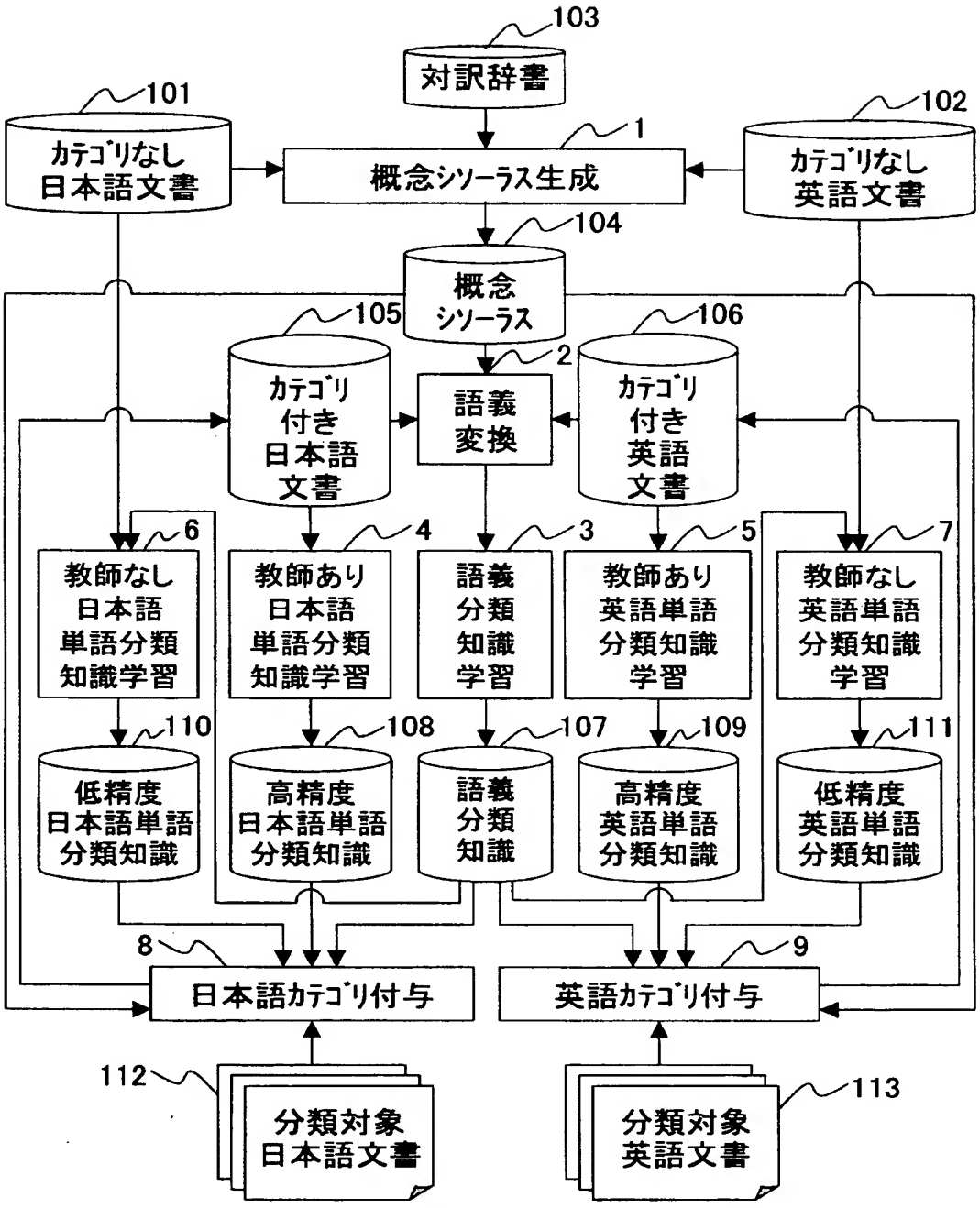
【図１１】単語ベクトルの例である。
【図１２】低精度英語単語分類知識学習処理の処理フローである。
【図１３】英語単語テーブルの例である。
【図１４】低精度英語単語分類知識学習処理の概念図である。
【図１５】教師なし学習による低精度英語単語分類知識学習処理の処理フローである。
【図１６】語義分類知識によるカテゴリ付与処理の概念図である。
【図１７】カテゴリ付与処理の処理フローである。
【図１８】インタラクティブなカテゴリ付与のための画面の表示例である。
【図１９】インタラクティブな分類知識修正のための画面の表示例である。

【符号の説明】

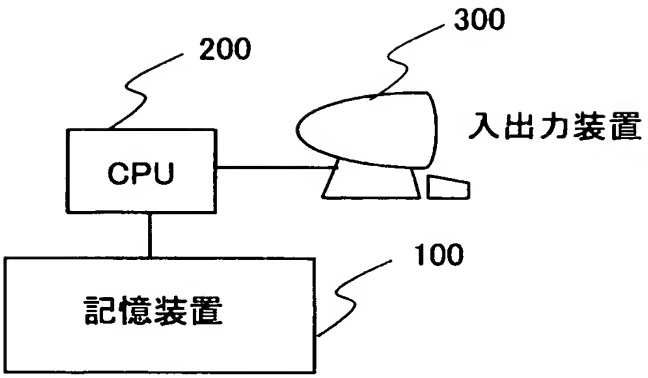
【００７６】

1　概念シソーラス生成プログラム
2　語義変換プログラム
3　語義分類知識学習プログラム
4　教師あり日本語単語分類知識学習プログラム
5　教師あり英語単語分類知識学習プログラム
6　教師なし日本語単語分類知識学習プログラム
7　教師なし英語単語分類知識学習プログラム
8　日本語カテゴリ付与プログラム
9　英語カテゴリ付与プログラム
100　記憶装置
200　ＣＰＵ
300　入出力装置
101　日本語文書
102　英語文書
103　対訳辞書
104　概念シソーラス
105　カテゴリ付き日本語文書
106　カテゴリ付き英語文書
107　語義分類知識
108　高精度日本語単語分類知識
109　高精度英語単語分類知識
110　低精度日本語単語分類知識
111　低精度英語単語分類知識
112　分類対象日本語文書
113　分類対象英語文書

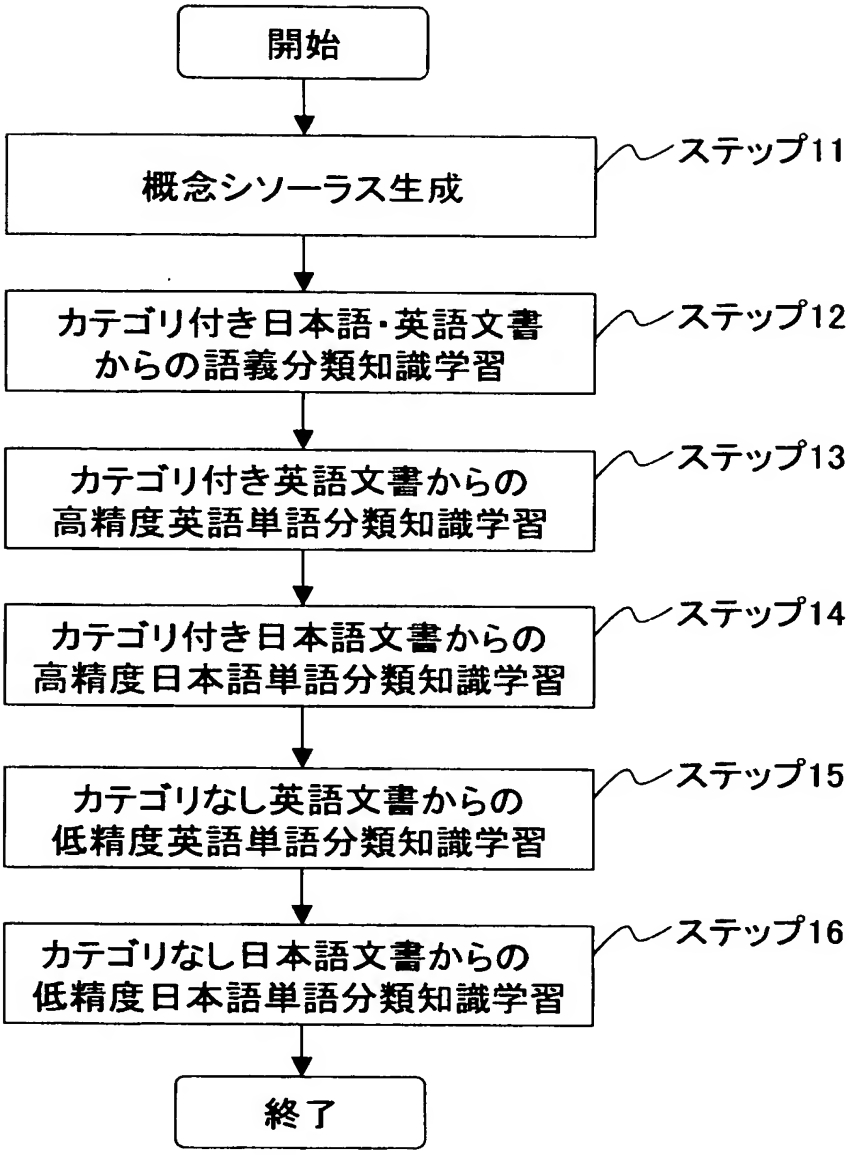
【書類名】図面
【図１】


103
対訳辞書
101
カテゴリなし
日本語文書
1
概念シソーラス生成
102
カテゴリなし
英語文書
104
概念
シソーラス
105
106
2
カテゴリ
付き
日本語
文書
語義
変換
カテゴリ
付き
英語
文書
6
4
3
5
7
教師なし
日本語
単語分類
知識学習
教師あり
日本語
単語分類
知識学習
語義
分類
知識
学習
教師あり
英語単語
分類知識
学習
教師なし
英語単語
分類知識
学習
110
108
107
109
111
低精度
日本語単語
分類知識
高精度
日本語単語
分類知識
語義
分類
知識
高精度
英語単語
分類知識
低精度
英語単語
分類知識
8
9
日本語カテゴリ付与
英語カテゴリ付与
112
113
分類対象
日本語文書
分類対象
英語文書

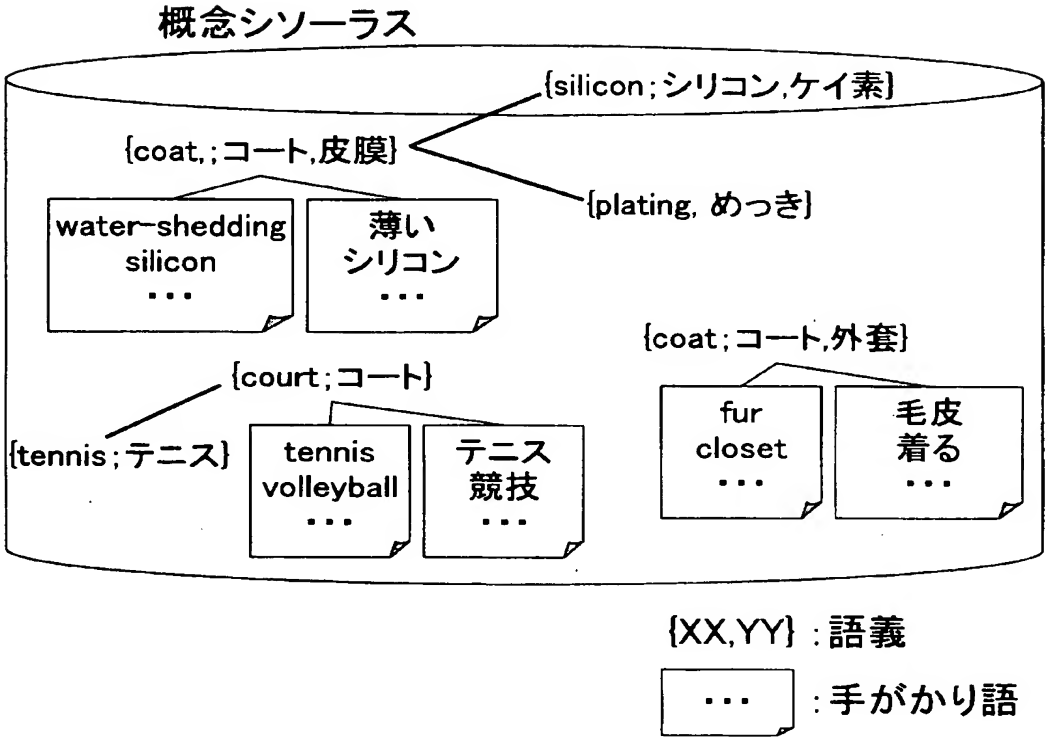
【図２】



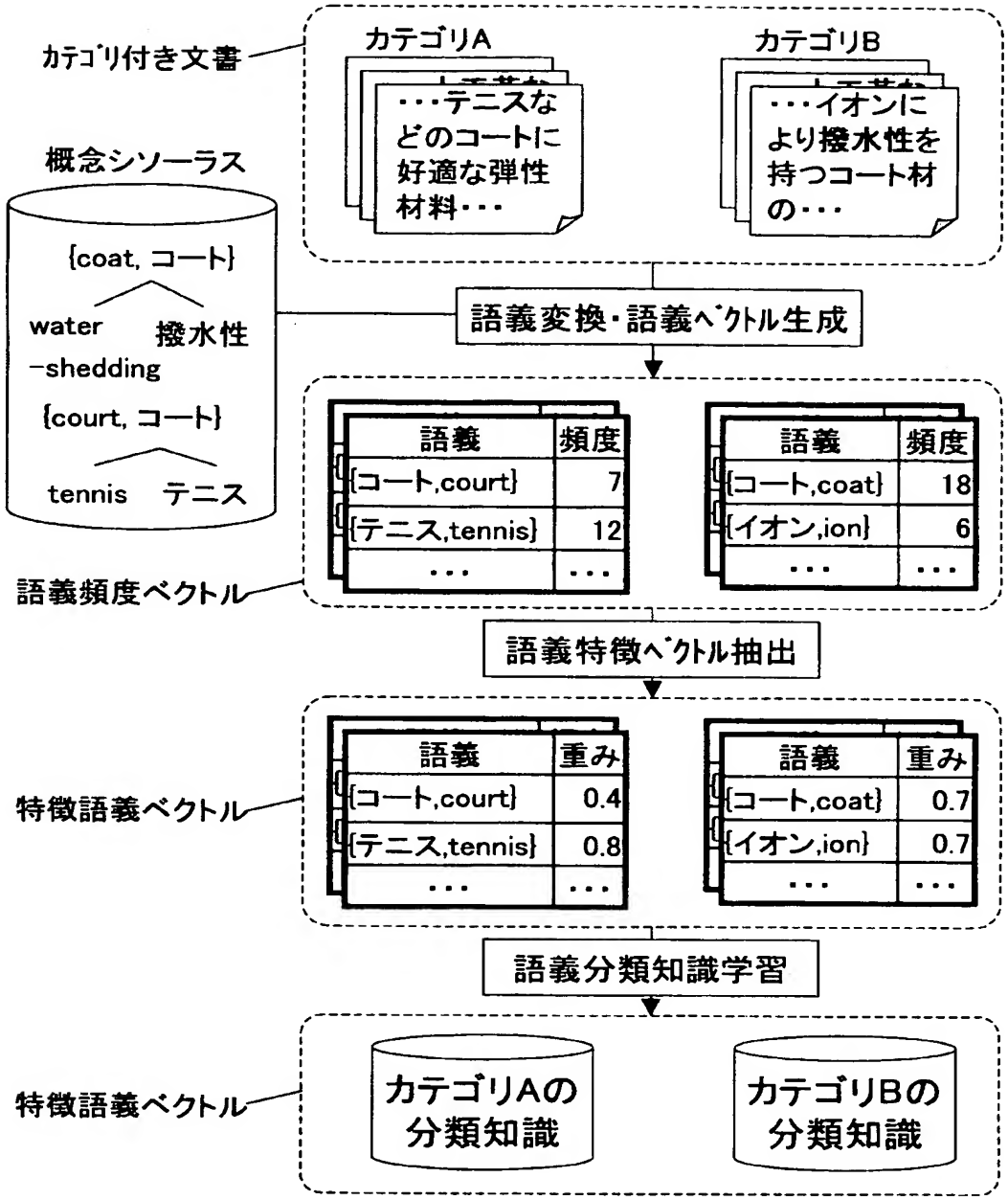
【図３】


開始
概念シソーラス生成
ステップ11
カテゴリ付き日本語・英語文書からの語義分類知識学習
ステップ12
カテゴリ付き英語文書からの高精度英語単語分類知識学習
ステップ13
カテゴリ付き日本語文書からの高精度日本語単語分類知識学習
ステップ14
カテゴリなし英語文書からの低精度英語単語分類知識学習
ステップ15
カテゴリなし日本語文書からの低精度日本語単語分類知識学習
ステップ16
終了

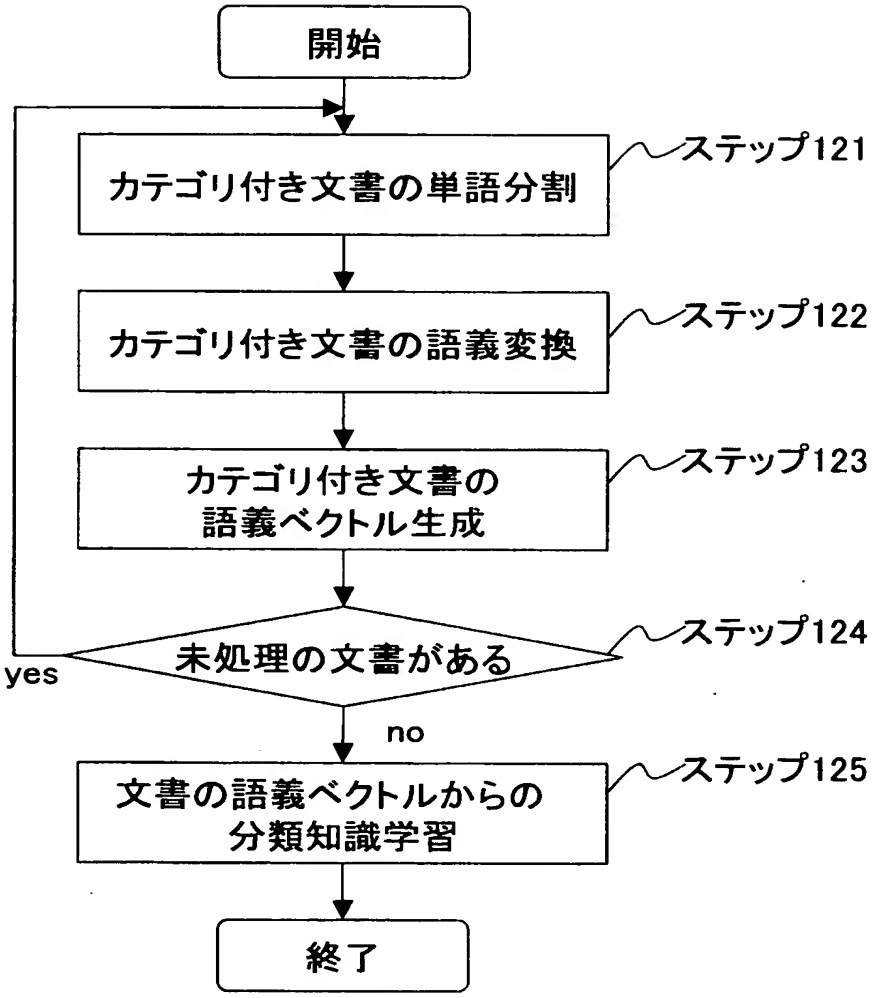
【図４】


概念シソーラス
{silicon;シリコン,ケイ素}
{coat,;コート,皮膜}
{plating, めっき}
water-shedding silicon
・・・
薄い シリコン
・・・
{coat;コート,外套}
{court;コート}
{tennis;テニス}
tennis volleyball
・・・
テニス 競技
・・・
fur closet
・・・
毛皮 着る
・・・
{XX,YY}：語義
・・・：手がかり語

【図５】


カテゴリ付き文書
カテゴリA
・・・テニスなどのコートに好適な弾性材料・・・
カテゴリB
・・・イオンにより撥水性を持つコート材の・・・
概念シソーラス
{coat, コート}
water
撥水性
-shedding
{court, コート}
tennis
テニス
語義変換・語義ベクトル生成
語義
頻度
{コート,court}
7
{テニス,tennis}
12
語義
頻度
{コート,coat}
18
{イオン,ion}
6
語義頻度ベクトル
語義特徴ベクトル抽出
語義
重み
{コート,court}
0.4
{テニス,tennis}
0.8
語義
重み
{コート,coat}
0.7
{イオン,ion}
0.7
特徴語義ベクトル
語義分類知識学習
カテゴリAの分類知識
カテゴリBの分類知識
特徴語義ベクトル

【図６】


開始
カテゴリ付き文書の単語分割
ステップ121
カテゴリ付き文書の語義変換
ステップ122
カテゴリ付き文書の
語義ベクトル生成
ステップ123
未処理の文書がある
ステップ124
yes
no
文書の語義ベクトルからの
分類知識学習
ステップ125
終了

【図７】

#	単語
1	シリコン
2	コート
3	で
4	被覆する
5	材料
6	は
7	撥水性
	・・・

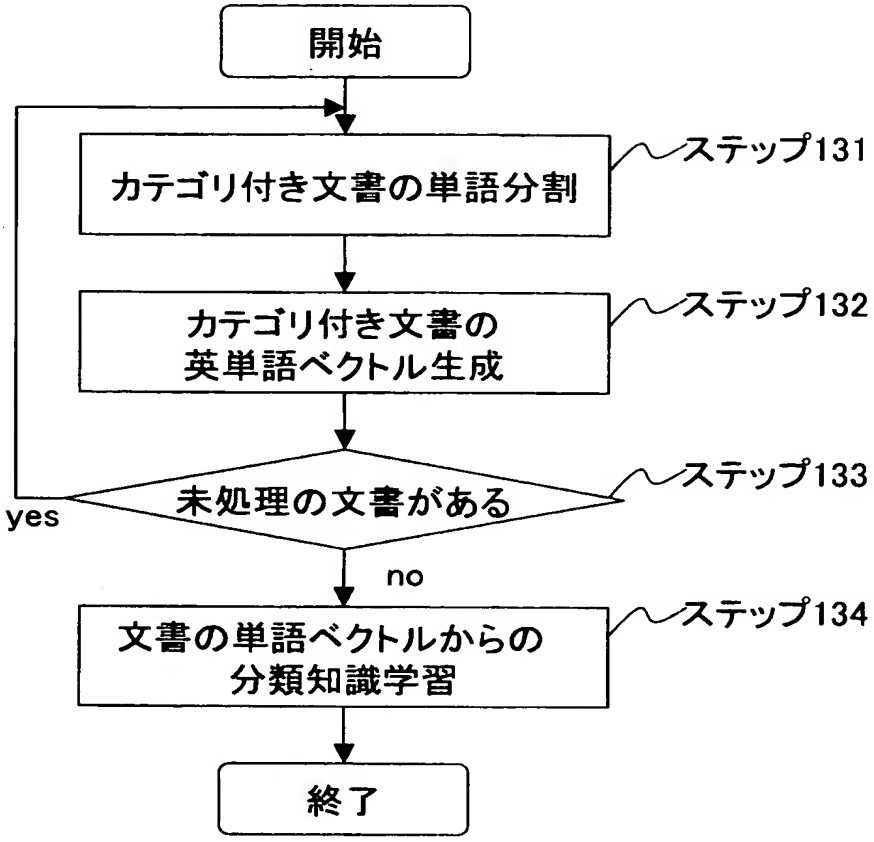
【図８】

#	単語	頻度
1	{coat,コート,皮膜}	7
2	{cover,カバーする,被覆する}	16
3	{material,材,材料}	21
4	{silicon,シリコン,ケイ素}	6
5	{water-shedding,撥水性}	12
	・・・	・・・

【図９】

#	単語	重み
1	{coat,コート,皮膜}	0.8
2	{cover,カバーする,被覆する}	0.4
3	{material,材,材料}	0.3
4	{silicon,シリコン,ケイ素}	0.6
5	{water-shedding,撥水性}	0.7
	・・・	・・・

【図１０】


開始
カテゴリ付き文書の単語分割
ステップ131
カテゴリ付き文書の
英単語ベクトル生成
ステップ132
未処理の文書がある
ステップ133
yes
no
文書の単語ベクトルからの
分類知識学習
ステップ134
終了

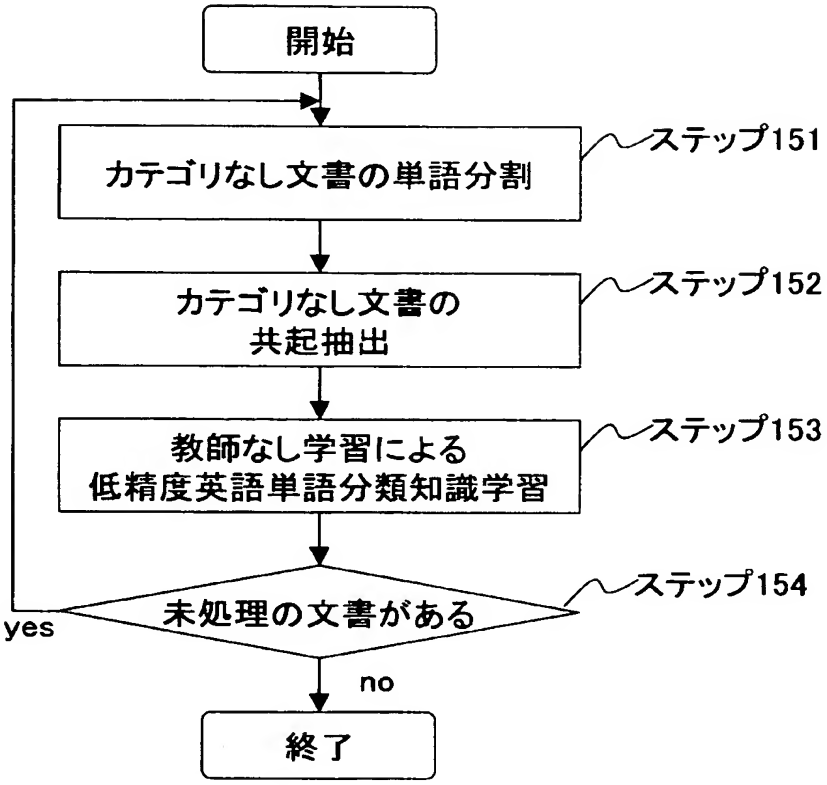
【図１１】

#	単語
1	cover
2	with
3	silicon
4	coat
5	water-shedding
6	material
7	is
	・・・

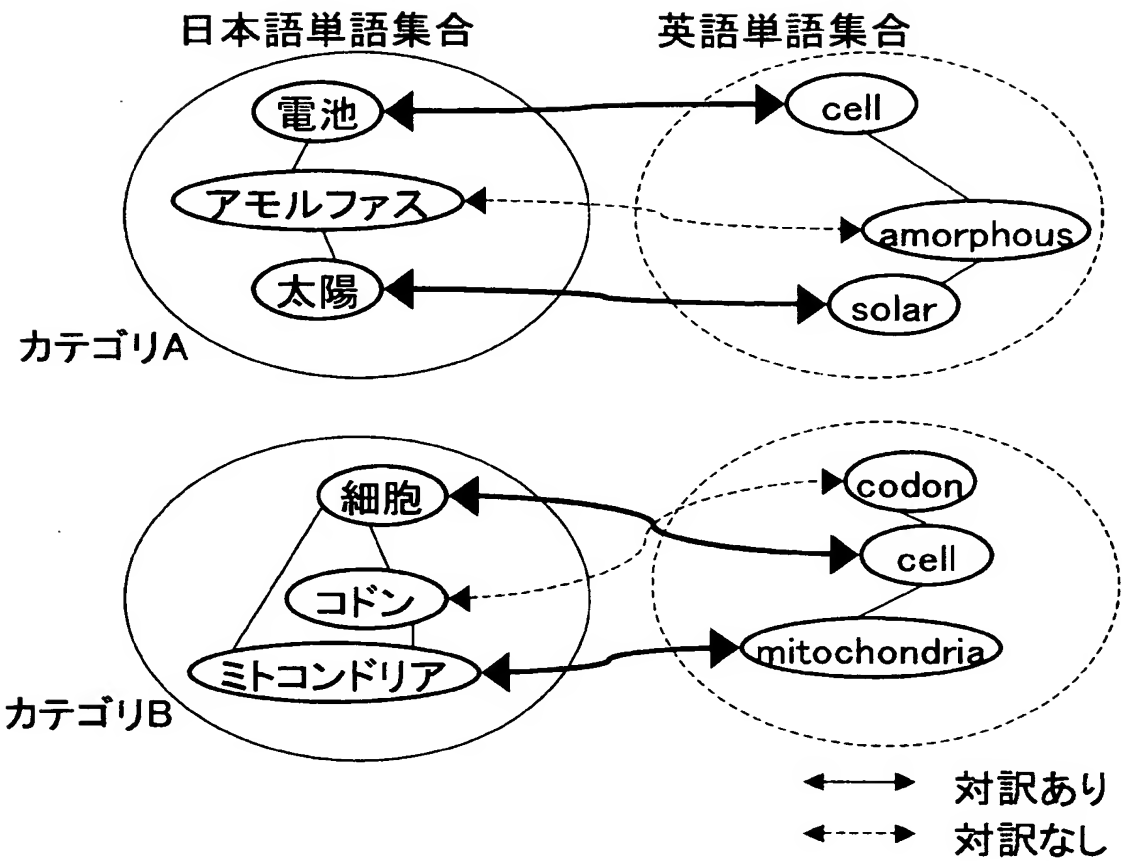
【図１２】

#	単語	頻度
1	coat	7
2	cover	16
3	material	21
4	silicon	6
5	water-shedding	12
	...	...

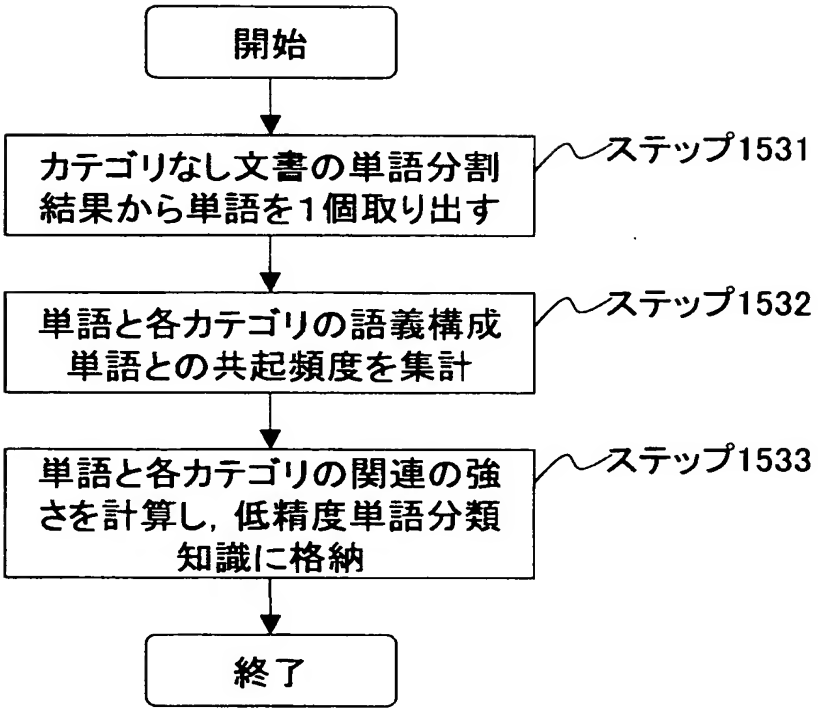
【図１３】


開始
カテゴリなし文書の単語分割
ステップ151
カテゴリなし文書の
共起抽出
ステップ152
教師なし学習による
低精度英語単語分類知識学習
ステップ153
未処理の文書がある
ステップ154
yes
no
終了

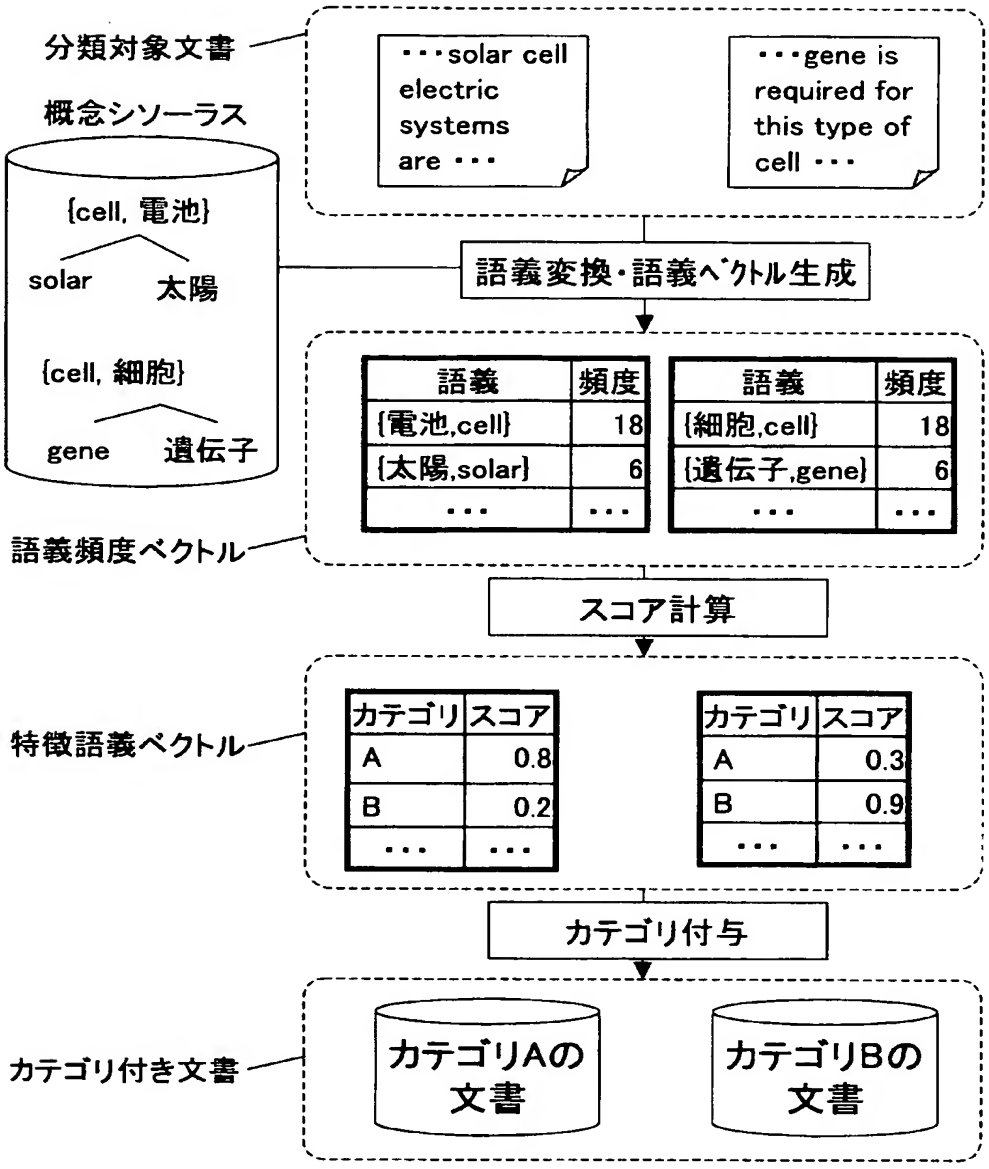
【図１４】


日本語単語集合
英語単語集合
電池
cell
アモルファス
amorphous
太陽
solar
カテゴリA
細胞
codon
cell
コドン
ミトコンドリア
mitochondria
カテゴリB
対訳あり
対訳なし

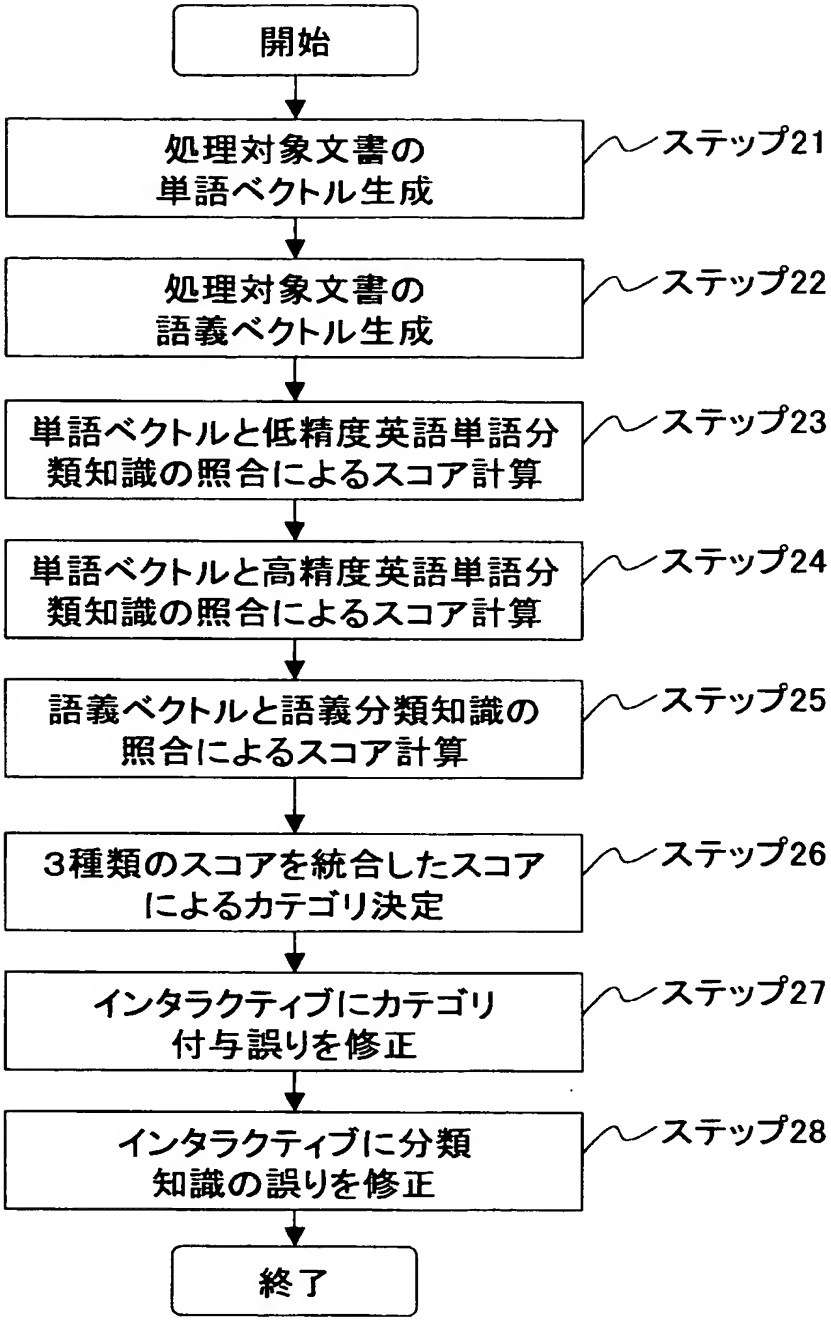
【図１５】


開始
カテゴリなし文書の単語分割結果から単語を1個取り出す
ステップ1531
単語と各カテゴリの語義構成単語との共起頻度を集計
ステップ1532
単語と各カテゴリの関連の強さを計算し，低精度単語分類知識に格納
ステップ1533
終了

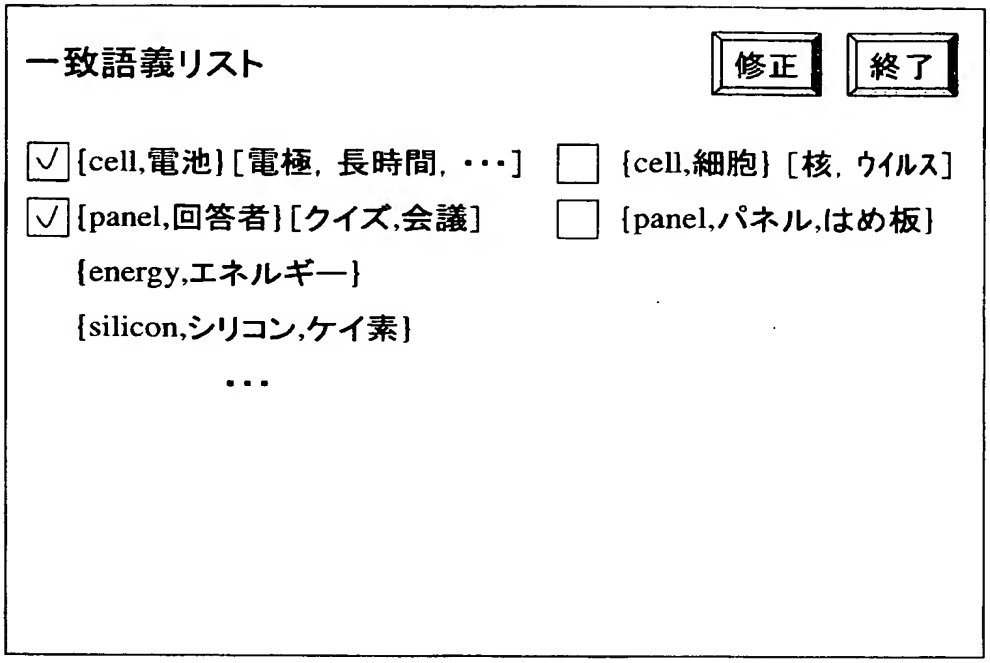
【図１６】


分類対象文書
概念シソーラス
・・・solar cell electric systems are ・・・
・・・gene is required for this type of cell ・・・
{cell, 電池}
solar
太陽
{cell, 細胞}
gene
遺伝子
語義変換・語義ベクトル生成
語義
頻度
{電池,cell}
18
{太陽,solar}
6
{細胞,cell}
18
{遺伝子,gene}
6
語義頻度ベクトル
スコア計算
カテゴリ
スコア
A
0.8
B
0.2
A
0.3
B
0.9
特徴語義ベクトル
カテゴリ付与
カテゴリAの文書
カテゴリBの文書
カテゴリ付き文書

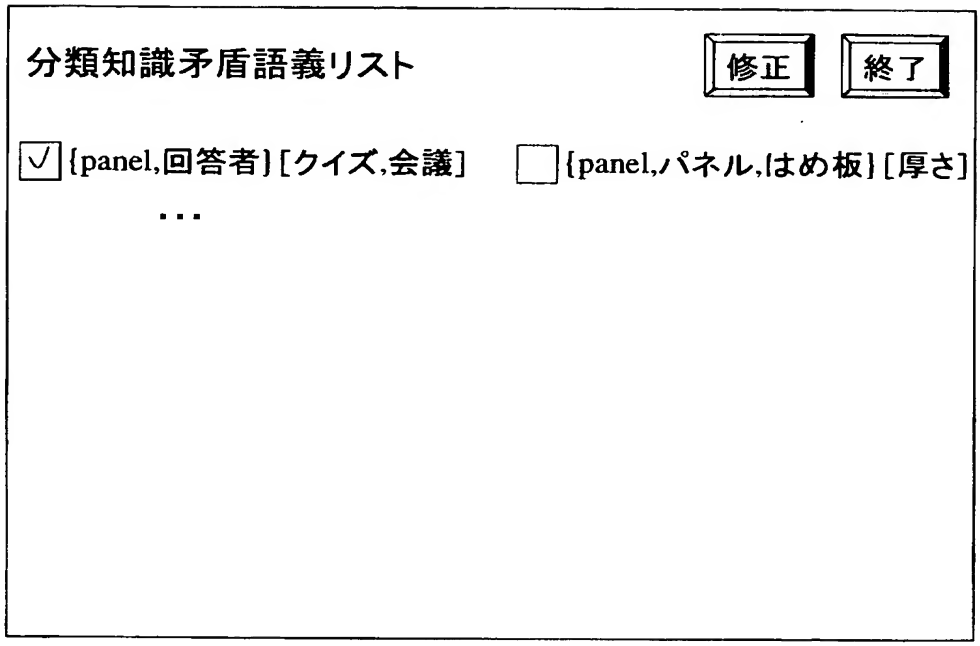
【図１７】


開始
処理対象文書の
単語ベクトル生成
ステップ21
処理対象文書の
語義ベクトル生成
ステップ22
単語ベクトルと低精度英語単語分
類知識の照合によるスコア計算
ステップ23
単語ベクトルと高精度英語単語分
類知識の照合によるスコア計算
ステップ24
語義ベクトルと語義分類知識の
照合によるスコア計算
ステップ25
3種類のスコアを統合したスコア
によるカテゴリ決定
ステップ26
インタラクティブにカテゴリ
付与誤りを修正
ステップ27
インタラクティブに分類
知識の誤りを修正
ステップ28
終了

【図１８】


一致語義リスト
修正
終了
{cell,電池}[電極，長時間，・・・]
{cell,細胞}[核，ウイルス]
{panel,回答者}[クイズ,会議]
{panel,パネル,はめ板}
{energy,エネルギー}
{silicon,シリコン,ケイ素}
・・・

【図１９】



【書類名】要約書
【目的】複数言語を対象とした文書分類方法を提供する。
【構成】カテゴリが付与されていない複数の言語の文書から語義および単語を語義に変換するための情報を抽出する手段と、カテゴリが付与されている文書から抽出した単語を語義に変換したのち語義レベルでの分類知識を学習する手段と、カテゴリが付与されている文書から単語レベルでの分類知識を学習する手段と、語義レベルの分類知識とカテゴリが付与されていない文書から抽出された単語の関連情報から単語レベルでの分類知識を学習する手段と、それぞれの分類知識を組み合わせてカテゴリの付与を行う手段、とを設ける。
【選択図】図１

特願２００３－３３８１７７

出願人履歴情報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所